Do your exposures need supervision?

Jenna Krall
Postdoctoral Fellow
Department of Biostatistics & Bioinformatics
Emory University
July 13 & 14, 2015

Coauthors: Howard H. Chang, Katherine M. Gass, W. Michael Caudle, Matthew J. Strickland

Acknowledgements: NIEHS (P30ES019776, K01ES019877, T32ES012160), EPA (R834799)

Overview of Methods

Unsupervised and supervised approaches

- ***Unsupervised***: Group predictors independent of the outcome
- ***Supervised***: Use the outcome to determine the best predictors

We used Principal Component Analysis (PCA) as our unsupervised approach and Classification and Regression Trees (C&RT) as our supervised approach.

Supervised

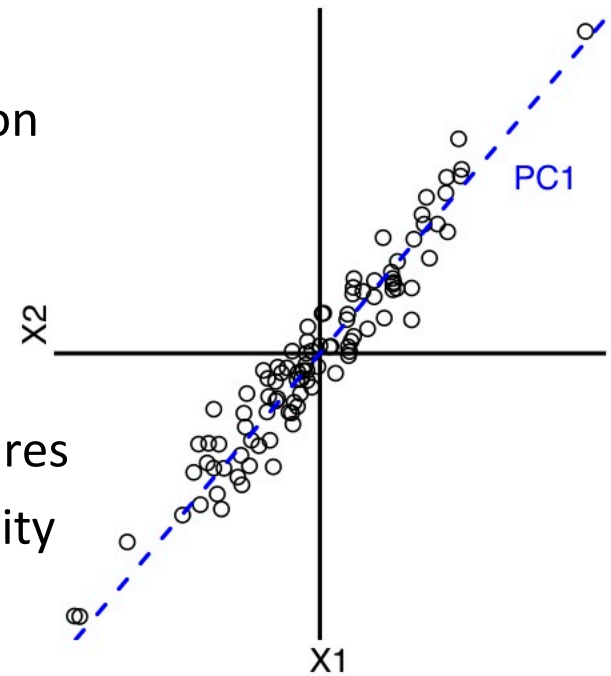
Unsupervised

Y	X1	X2	X3
-0.14	-1.01	-0.92	0.54
0.77	0.39	-1.01	-0.06
1.82	0.38	-0.95	0.48
2.35	0.93	0.23	0.33

Principal Component Analysis (PCA)

Overview for PCA approach:

- Groups exposures together based on correlation
- Decreases multicollinearity in multivariate regression models



Methods:

1. Apply PCA to the correlation matrix of exposures
2. Use varimax rotation to improve interpretability
3. Regress outcome on rotated principal components (rPCs)

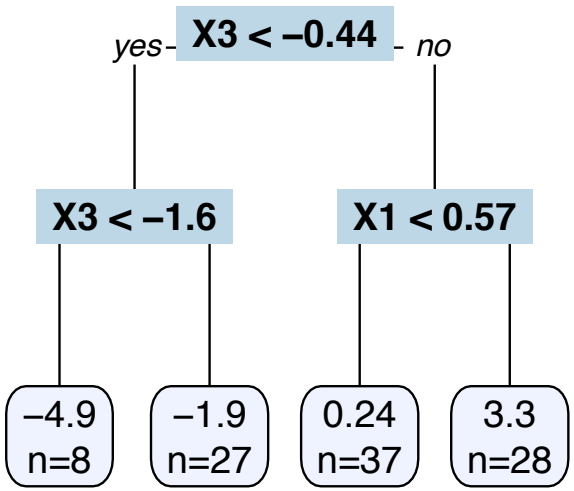
Assumptions:

- Exposures are approximately multivariate normal

Classification and Regression Trees (C&RT)

Overview for C&RT:

- Uses dichotomous splits of the exposures to predict the outcome
- Chooses most important exposures recursively



Methods:

1. Regress out effects of confounders from the outcome and exposures
2. Apply C&RT to residuals
3. Prune final tree based on cross validation

Assumptions:

- Fewer assumptions than many models

Final model

PCA approach

- Identified 6 rotated PCs (rotPCs) that explained 88.7% of the variability in the exposures
- Final adjusted model (*g* is natural spline with 3 df)

$$E(Y) = \beta_0 + \sum_{k=1}^{6} \beta_k \times rotPC_k + \gamma_1 Z_1 + \gamma_2 Z_3 + g(\phi; Z_2)$$

C&RT

- Confounder model for outcome *Y* and exposures *X1-X14*

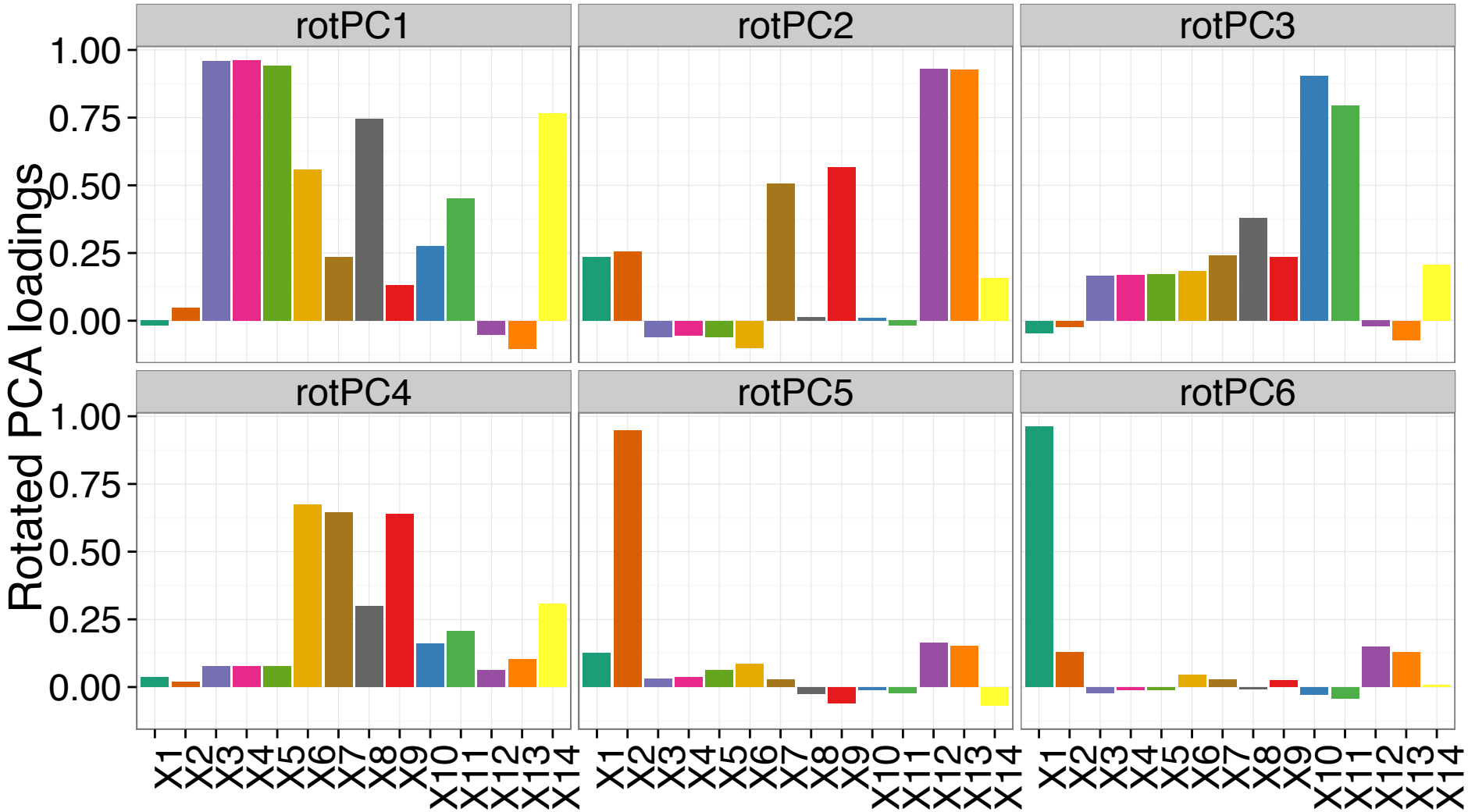
$$E(Y) = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_3 + g(\eta; Z_2)$$

- Use all exposures in C&RT model

Software and Code Used

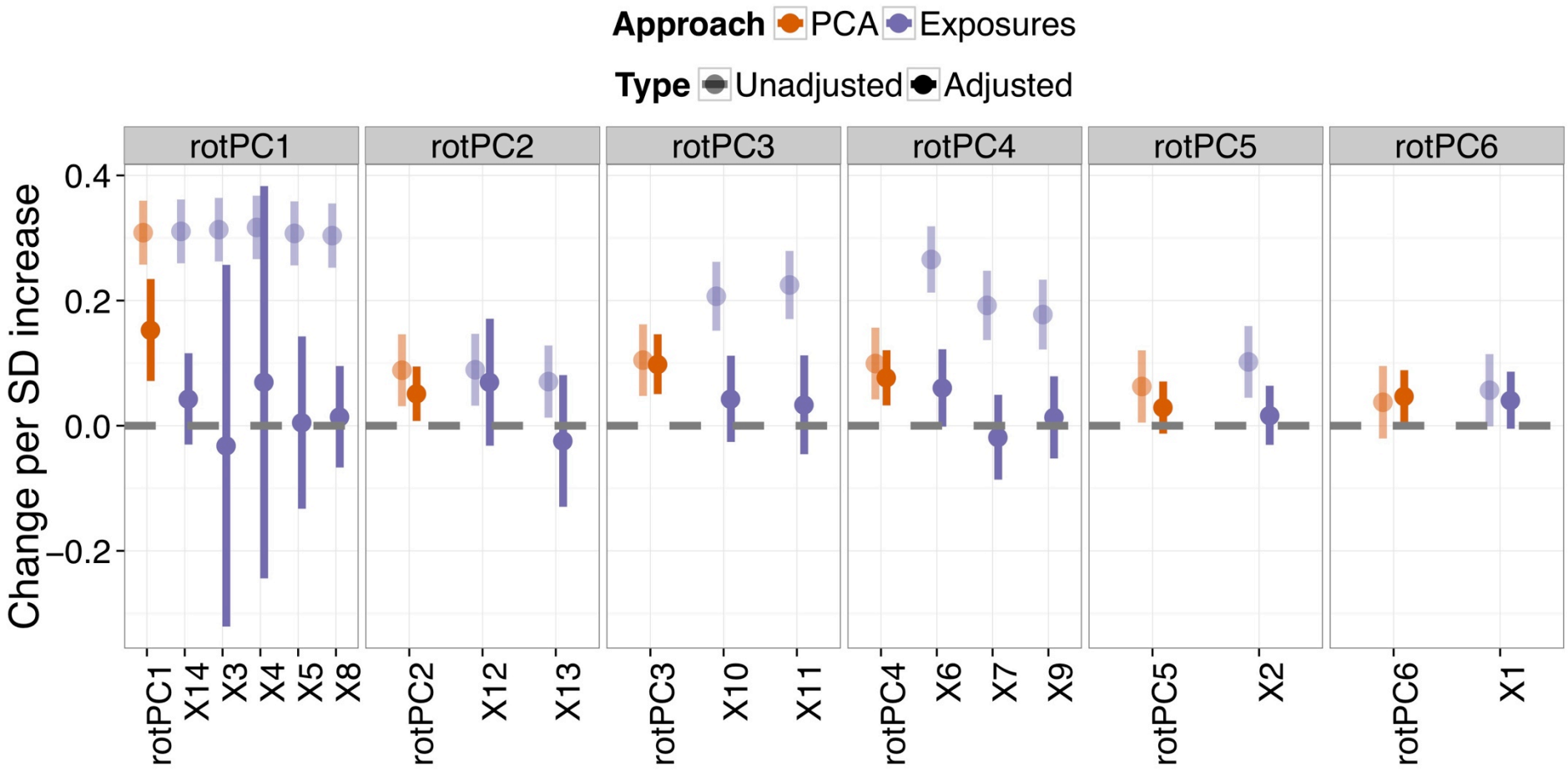
- R version 3.1
- Existing functions
- PCA approach
  - `principal() in psych`
  - version 1.5, Revelle (2015)
- C&RT
  - `rpart() in rpart`
  - Version 4.1 Therneau, et al. (2015)
- Full code available
  - `github.com/kralljr/niehs-epistats`

Varimax rotated principal component (*rotPC*) loadings



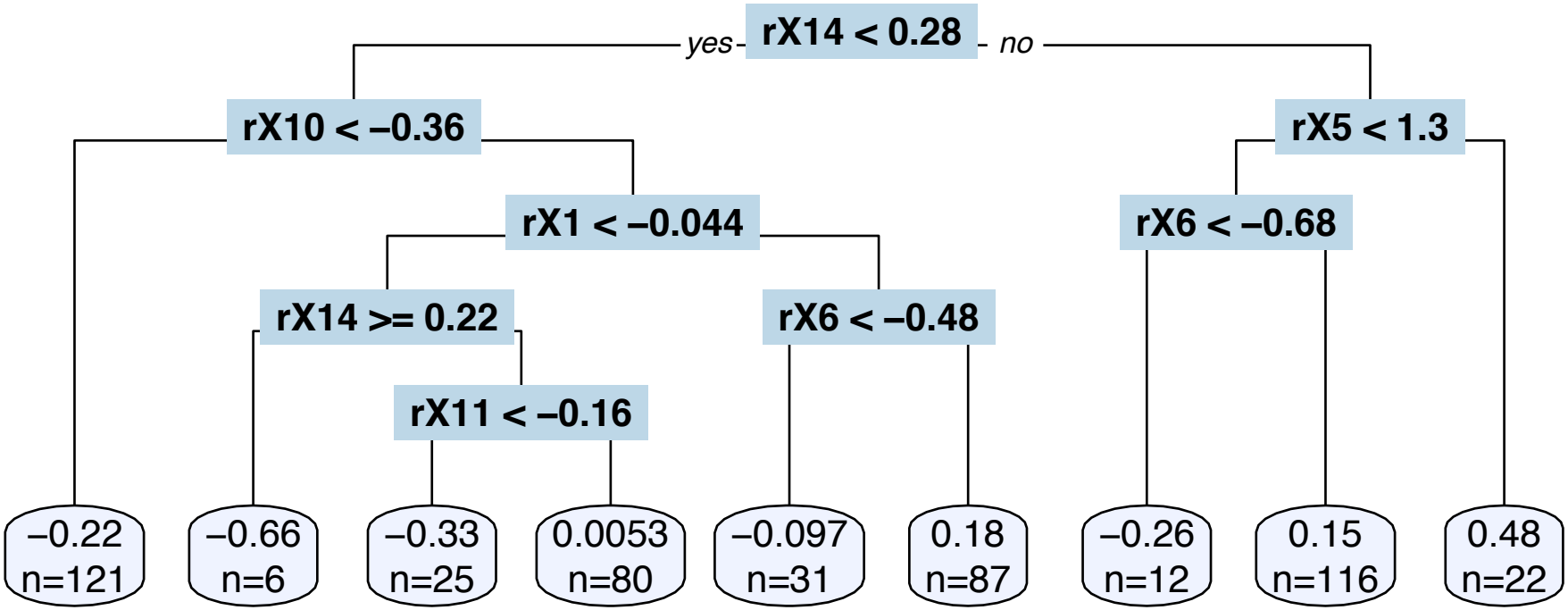
Estimated coefficients using the PCA approach and the exposures directly.

- Adjusted models control for other predictors and *Z1-Z3*

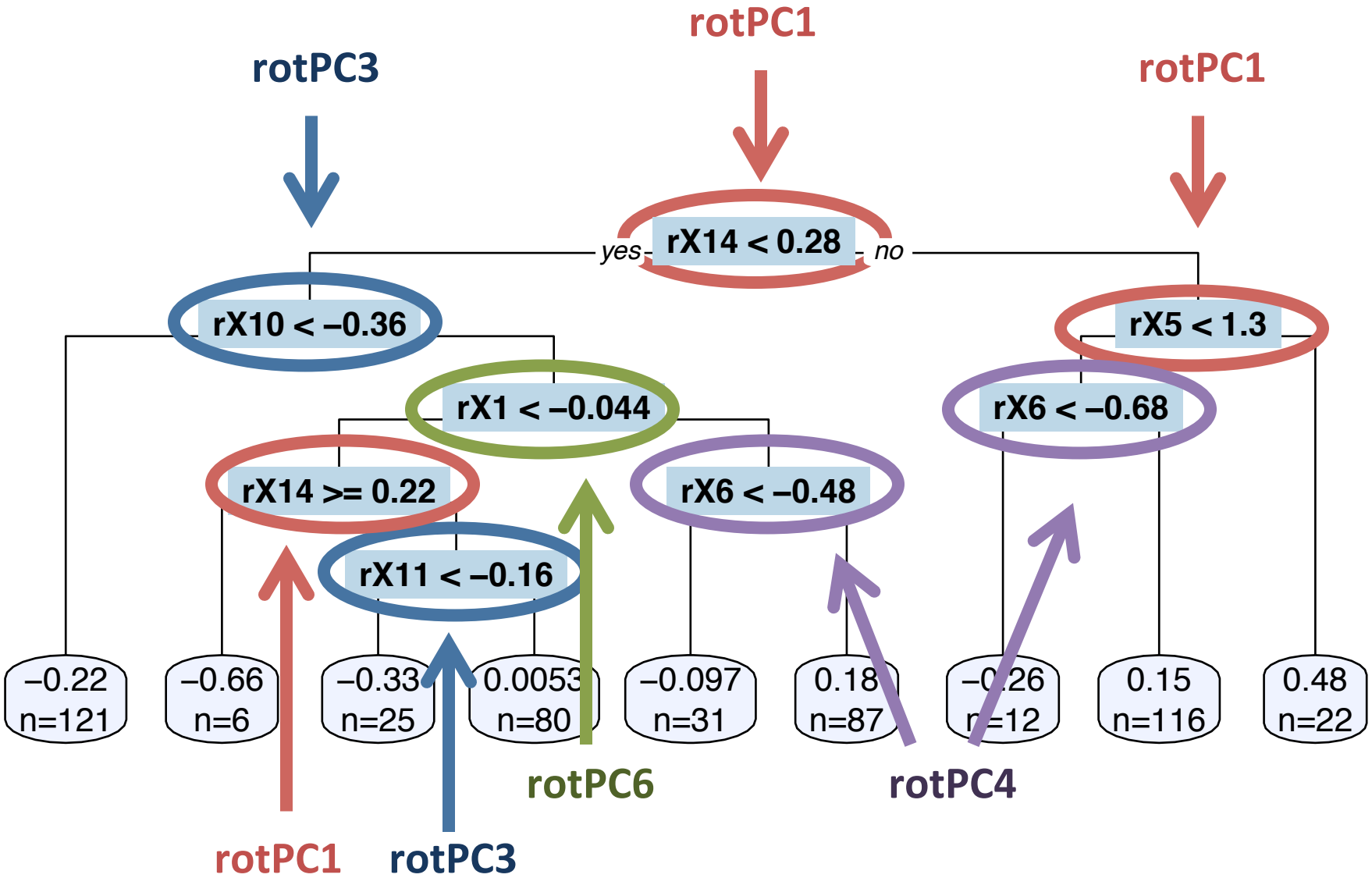


Pruned regression tree using C&RT

- No exposures were found to significantly predict the outcome using cross validation
- We highlight the most important predictors.



None of the top exposures correspond to *rotPC2* or *rotPC5*



Results

1. Which exposures contributed to the outcome?
   - PCA: Positive associations for *rotPC1* (*X3-X5, X8, X14*), *rotPC2* (*X12-X13*), *rotPC3* (*X10-X11*), *rotPC4* (*X6-X7, X9*), and *rotPC6 (X1)*.
   - C&RT: Did not find any exposures were predictive of the outcome. First splits occurred on (in order) *X14, X10, X5, X1, X6, X11*.
2. How much did the exposures contribute to the outcome?
   - PCA: Final model

$$E(Y) = 3.8 + \underline{0.15 \times rotPC1} + 0.05 \times rotPC2 + \underline{0.10 \times rotPC3} + \underline{0.08 \times rotPC4} + 0.03 \times rotPC5 + 0.05 \times rotPC6 + confounders$$

     - The final model explained 51.2% of the variability in *Y*.
     - The likelihood ratio test demonstrated that the final model fit the data better than a model using the confounders alone (*p<0.01)*.
   - C&RT: Exposures did not significantly contribute to the outcome using cross validation.

Results

3. Was there evidence of interactions?
   - PCA: We found some evidence of interaction between *rotPC1* and *rotPC3*.
   - C&RT: Difficult to evaluate interactions.
4. What was the effect of joint exposure to the mixture?
   - PCA: Final model

$$E(Y) = 3.8 + 0.15 \times rotPC1 + 0.05 \times rotPC2 + 0.10 \times rotPC3 + \\ 0.08 \times rotPC4 + 0.03 \times rotPC5 + 0.05 \times rotPC6 + confounders$$

   - C&RT: Un-pruned tree indicates that increased exposures were generally associated with increased Y.

Reflection

Did PCA get the right answer?

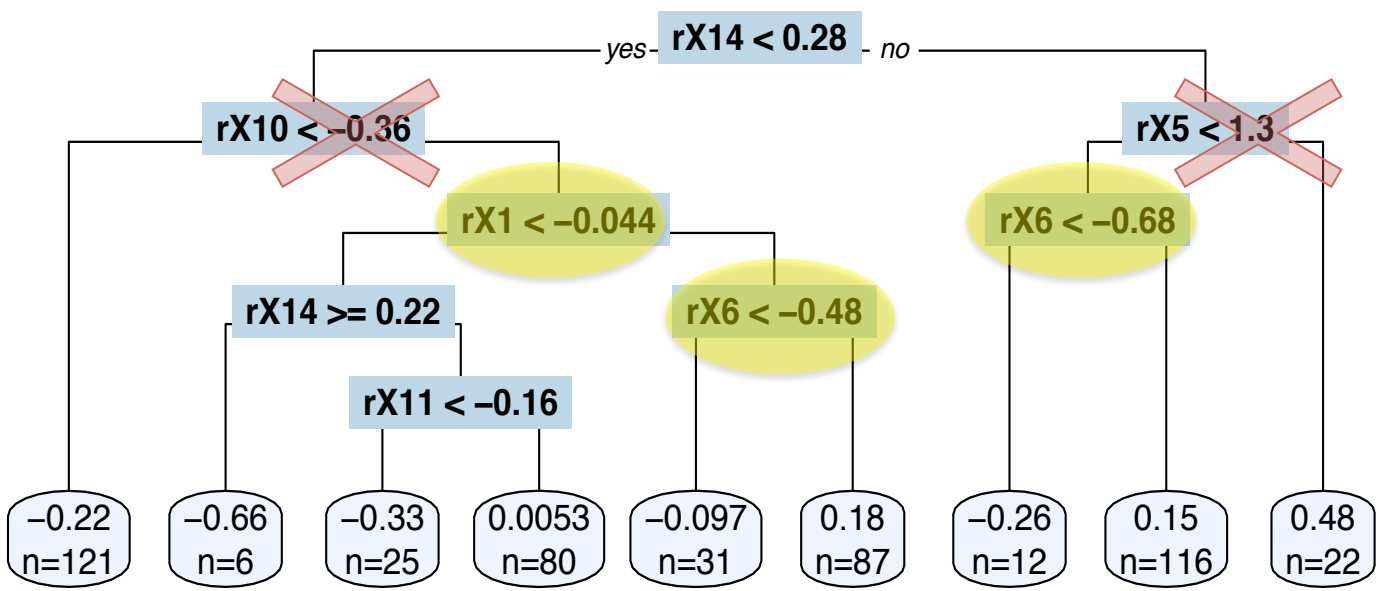
- True exposures all contributed to the rPC groups most associated with Y
- Could not determine which exposures were most important using PCA
- Varimax rotation yields results closest to truth

PCA approach		**Truth**	**Z3 = 0**	**Z3 = 1**
rotPC1	0.15	X14	0.10	0.10
		X4	0.05	0.05
rotPC2	0.05	X12	0.50	*0*
rotPC3	0.10	X11	0.10	0.10
rotPC4	0.08	X6	0.10	*0*
rotPC5	*0.03*	X2	*0*	*0*
rotPC6	0.05	X1	*0*	0.01

Reflection

Did C&RT get the right answer?

- No splits were predictive of the outcome
- Associations were mostly in the correct direction
- X4 and X12 not in the top splits



Summary

Strengths and weaknesses

- Both approaches are easy to apply and are implemented in most statistical packages.
- A lack of consistency between unsupervised and supervised approaches may indicate uncertainty in the results.
- Each approach has its own strengths and limitations.
  - PCA: Groups correlated exposures together, but resulting PCs may not be very interpretable.
  - C&RT: Fewer assumptions, but does not find exposures predictive of the outcome in the presence of strong confounding. It is difficult to add complexity to the model.

Next steps

- Compare PCA approach and C&RT to other supervised and unsupervised approaches.